Flexible geostatistical modelling and risk assessment analysis of lead concentration levels of residential soil in the Coeur D'Alene River Basin

Dae-Jin Lee*[1] and Peter Toscas[1]

[1]Commonwealth Scientific and Industrial Organization, Computational Informatics

December 17, 2014

Abstract

Soil heavy metals pollution is an urgent problem worldwide. Understanding the spatial distribution of pollutants is critical for environmental management and decision-making. Children and adults are still routinely exposed to very high levels of heavy metals contaminants in some countries, particularly in regions with a long mining history. In this paper, we analyze lead concentration levels from residential soil samples in the Coeur D'Alene River Basin (CDRB) in the United States. The aim of this paper is to estimate the spatial distribution of the lead concentration levels that may affect exposed humans. Geographic coordinates were compiled for a total of 781 residential addresses and 1075 mine-related sites (e.g. mine tailings, rock dumps, mine wastes, etc.) surrounding the properties. The lead concentration levels analyzed in the study are in general far from uniform within a residential property and measured levels can differ greatly from one residential address to a nearby address. We consider a unified approach to model the lead concentration levels by means of penalized regression splines and tensor product smooths, using generalized linear models as a building block. We also use this approach to perform a risk assessment spatial analysis of map hot spots for lead based on the action levels defined by the U.S. Environmental Protection Agency.

1 Introduction

The Coeur D'Alene River Basin extends from the Idaho-Montana border on its eastern side to the Idaho-Washington border on its western side. It covers around 6,000 square kilometres in Shoshone and Kootenai Counties in northern Idaho. The Upper Basin contains 11 residential cities or unincorporated areas, about half of which are located

*Corresponding author. CSIRO CCI Private Bag 33, Clayton, VIC 3169, Australia Tel: +61 3 9545 8071 | Fax: +61 3 9545 8080
e-mail address: dae-jin.lee@csiro.au

within the Bunker Hill Superfund Site (BHSS), a historic mining and smelting district. In 1983, and subsequently in 1998, parts of the area were declared Superfund sites by the US Environmental Protection Agency (EPA). The smelter closed in 1981. Since the closure, an agreement between the Idaho Department of Environmental Quality (IDEQ) and the U.S. Environmental Protection Agency (EPA) has resulted in remedial actions with respect to reducing soil and dust levels. The aim is to identify potential human risks from lead (Pb) contamination in residential soil (see U.S. Environmental Protection Agency, 2002; Elias and Gulson, 2003, for details).

In 1985, a comprehensive plan of intervention and risk reduction was established to minimize lead absorption during the remedial investigation and cleanup phases of the Superfund project. Two major health response actions were implemented, combining in-home intervention, public awareness efforts, and targeted remedial activities: the Lead Health Intervention Program (LHIP) and the Residential Soil Cleanup (RSC). The LHIP involved and annual door-to-door blood lead surveys, nursing follow-up, and public education in schools, for parents and health care providers. However, biological data from blood lead surveys of the LHIP are not available due to confidentiality issues, so we only considered residential soil samples in this study. Lindern et al. (2003) identified some potential bias due to the decreasing degree of participation and parental reasons for refusing to be taken from their children.

Decisions for the Coeur D'Alene Basin (U.S. Environmental Protection Agency, 2002) as well as the Human Health Risk Assessment (TerraGraphics, 2003) provide excellent background and historical information on sampling and clean-up activities that have occurred in the Basin. For more than 100 years, the Coeur D'Alene Basin was a major producer of silver, lead, zinc and other metals. These activities have resulted in widespread heavy metals contamination. Mining related activities generate tailings, waste rock, sediments, and smelter emissions that contain elevated levels of metals. Most of the tailings were transported downstream, particularly during high flow events, and deposited as sediments in the bed, floodplains, and lateral lakes of the Upper and Lower Basin.

Further, tailing material was also dispersed via other means such as the use of railroad cars to tranport fill material for construction of roads, railroads and buildings, which resulted mining waste accumulating along rail road lines. Mining waste was also dispersed as airborne dust.

The quantities of tailings discharged to the Coeur D'Alene River Basin constitute a substantial amount of materials (U.S. Environmental Protection Agency, 2002). The amount of tailings, tailing-contaminated sediments and their metal content remaining in the Coeur D'Alene River is very difficult to determine and constitutes a major source of metals contamination in the Basin (TerraGraphics, 2003).

In this paper, we use residential soil sample data collected from surveys conducted during April to October of 2003. We focus on lead (Pb). At high concentrations, lead is a potentially toxic element to humans and other forms of life. There are two major sources of lead contamination: 1) lead-based paint where contamination may occur when paint chips from old buildings and mix with the soil; and, 2) lead from car emissions. The most serious source of exposure to soil lead is through direct ingestion (eating) of contaminated soil or dust. Preschool-age children and pregnant wowen are the most vulnerable segment of population for exposures to soil lead. People ingest lead in water, food, soil, and dust. In out study, the target population is residential property located within the boundaries of the CDRB with particular interest in homes with young chil-

dren and/or pregnant women. Samples were collected at the homes of residents that agreed to participate in the sampling effort, if the resident/renter refused to participate, solicitation continued at the next house. Soil was sampled in areas such as driveways, gardens, parking areas, play areas, yards and other areas such as sidewalks, areas under trees or near painted surfaces, following a protocol previously used by the State of Idaho in sampling residential properties in the BHSS and the rest of the Coeur D'Alene Basin (see TerraGraphics, 2003, for further details). Residential properties were sampled to identify those residences eligible for remediation or greening action based on lead concentrations in the soil. Removing the sources of heavy metal exposures it is hoped will reduce potential human health risks, particularly for young children and pregnant women. It is important to notice that the sampling protocol, data collection and assessment activity was undertaken with no statistical sampling design methodology.

In this paper we propose a retrospective analysis of the data collected in the residential addresses that agreed to collaborate in the 2003 study. The aim is to characterize a complex region in order to map the Pb concentration in soil in those residential areas near mining, smelting industrial complexes and tailings deposits. We propose a framework based on the use of flexible smoothing techniques in order to: (i) estimate a spatial surface that describes the spatial variability in the residential area of interest; (ii) incorporate the information of the mine tailings as a main source of heavy metal contamination and (iii) quantify the risk assessment of heavy metals relative to threshold values defined by the established action levels for Pb, that may be of practical importance at sampled and unsampled sites, and to quantify the risk of exceeding the established action levels. In the next section we provide details about the data considered in the study for which we apply the methodology. In section 3 we present the methodology and model formulation for Pb concentration levels in residential soil samples. In section 4 we reformulate the model to perform a geostatistical risk assessment to spatially locate exposure zones based on the action levels for remediation described in the protocols of the (U.S. Environmental Protection Agency, 2002), thus highlighting critical areas that may be used for targeted intervention. We end the paper with a discussion.

2 The data

The data consists of 781 unique residential addresses in different towns in the Upper Basin (e.g. Osburn, Wallace, Cataldo, Kellogg, Silverton, or Mullan among others). We consider Pb concentration levels in mg/kg units. The geographical coordinates were matched with the addresses recorded in the 2003 database. The locations of the residential properties used in this study are shown in Figure 2.1. The figure also shows the locations of the 1,075 mine-related sites surrounding the residential properties (which include tailings and tailing ponds, mine adits, rock dumps, mining materials used for construction, or mine tunnels). For each residential property up to eight different sample locations were chosen (Driveway, Garden, Garage, Parking area, Play area, Right-of-Way, Yard and other samples), at four different sample depth intervals (in inches): A (0-1), B (1-6), C (6-12) and D (12-18 in). The maximum number of combinations of both factors would be 32. As many properties only have a yard, driveway or garden areas to sample, bringing the average number of samples per home down was only 15. As a result the design is very unbalanced with only a small number of samples in some sample locations and sample depths. Table 2.1 shows the number of observations by sample location and depth. Further details about the data, sampling protocols and remediation

activities can be found in (TerraGraphics, 2003).

Table 2.1: Number of Pb samples used in the study by sample location and depth.

Pb	A (0-1 in)	B (1-6 in)	C (6-12 in)	D (12-18 in)
Driveway Sample	334	335	335	333
Garden Sample	250	251	251	247
Garage	32	32	30	30
Other Sample	364	356	358	355
Parking	207	209	213	202
Play Area Sample	12	11	12	11
Right-of-Way	919	916	921	907
Yard Sample	1484	1486	1490	1469

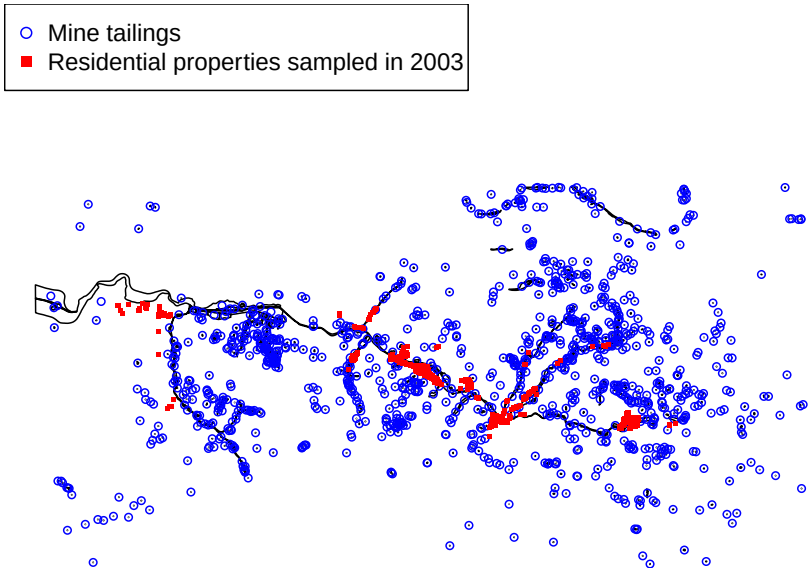


Figure 2.1: Residential properties used in this study. Black lines corresponds to main roads.

3 Spatial modelling of Lead concentration levels

Geostatistics has been popularly applied for investigating and mapping soil pollution by heavy metals (Goovaerts, 1997), however none of the previous studies of the CDRB have considered a geostatistical approach. It is important to remark that the surveys providing the data were not specifically designed to accommodate statistical techniques, hence caution should be exercised (Lindern et al., 2003). Samples were not randomly chosen and were targeted at high-risk homes or those that agreed to participate in the sampling effort. Because different remedial strategies were undertaken in different communities in different years, soil exposure reductions vary by neighbourhoods and community-wide environment. There are also a variety of factors contributing to the residential

property Pb levels that can make it more difficult to assess geographical patterns in exposures. For example the house age and the use of lead-based paints for houses built before 1960 when the use of lead-based paints were banned (Spalinger et al., 2007).

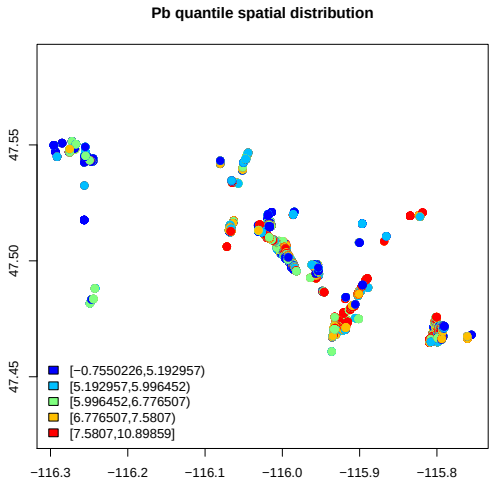


Figure 3.1: Spatial visualization of Pb concentration levels quantiles in logarithmic scale.

Classic geostatistical techniques, such as kriging (Cressie, 1993) are based on the assumption that data are a realization of a process with a second-order stationary multivariate distribution. The assumption of second-order stationarity means that the covariance function exists and the variogram is therefore bounded. Previous analysis of the CDRB showed that heavy metals contamination of soil is heterogeneously distributed and consequently, the level of contamination can differ greatly in short distances (Elias and Gulson, 2003; Lindern et al., 2003). In fact, lead levels are far from uniform within a residential property, sample location and sample depth. Figure 3.1 shows the quantiles of Pb concentration levels (log scale) in the residential properties considered. This spatial visualization of the data shows the existence of exceptionally high heterogeneity that complicates the variogram analysis and would violate the classic geostatistics theory assumptions (such as stationarity and isotropy). The correct specification of the spatial covariance and its parameters might be of importance when prediction is the aim, however, in this paper, due to the heterogeneity of the lead concentration levels, we are interested in the assessing the mean levels of Pb concentrations in the whole CDRB area. We use a semi-parametric regression modelling approach where the bivariate spatial surface is modelled by means of low-rank tensor products of spline basis functions which are not constrained to the selection of a spatial covariance matrix or make other strong assumptions (Eilers and Marx, 1996; Currie et al., 2006; Wood, 2006b).

A number of authors have compared kriging and non-parametric regression techniques in the statistics literature (see for instance Laslett (1994) or Wahba (1990) among others). Penalized regression splines have become a very popular technique for bivariate smoothing. Indeed, kriging can be viewed as a spline type model, as in theory a kriging estimate is identical to a thin plate spline for a particular generalized covariance function. Kammann and Wand (2003) combine the ideas of geostatistics and smooth

modelling in an additive framework (Hastie and Tibshirani, 1990) and called it geoadditive models.

3.1 Spatial data modelling with low-rank smoothers

Consider geostatistical data of the form $(\boldsymbol{s}_i, \boldsymbol{y}_i)$, for $i = 1, ...n$, where $\boldsymbol{y}_i$ is the continuous outcome variable and $\boldsymbol{s}_i \in \mathbb{R}^2$ represent the spatial locations. A non-parametric model for the data is given by:

$$\boldsymbol{y}_i = f(\boldsymbol{s}_i) + \boldsymbol{\epsilon}_i, \quad 1 \leq i \leq n, \tag{3.1}$$

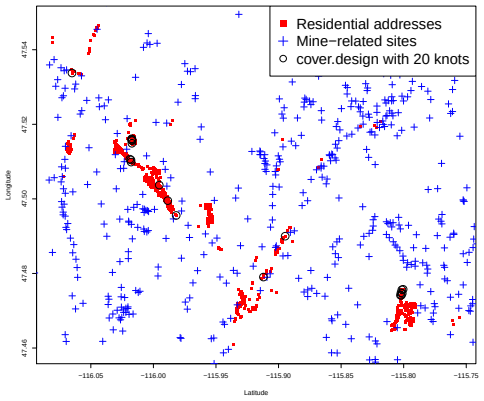
where $f(\cdot)$ is an unknown smooth bivariate function of the locations $\boldsymbol{s}_i = (Lon_i, Lat_i)'$. We assume that the vector of regression errors $\boldsymbol{\epsilon}$ is i.i.d. normal, i.e. $\text{Cov}(\boldsymbol{\epsilon}) \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. The problem of modelling the function $f(\cdot)$ has many statistical solutions. While kriging assumes that the regression function is a polynomial and the errors are second-order intrinsically stationary with a parametric correlation structure depending on the distance (see Cressie, 1993). A spline-based basis representation for the function $f(\cdot)$ might be written as $f(\boldsymbol{s}) = \sum_{j=1}^{m} \alpha_j \phi(\boldsymbol{s})$ where α_j are a set of coefficients and $\{\phi_j(\boldsymbol{s}), j = 1, 2, ..., m\}$ are spline basis where in general $m < n$. A very convenient formulation of model in Eq. (3.1) is as a linear mixed model. Mixed model representations in non-parametric regression have been used by many researchers in recent years (e.g. Wang (1998); Brumback and Rice (1998); Lin and Zhang (1999); Verbyla et al. (1999)). Model (3.1) can formulated as a mixed model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{G}), \tag{3.2}$$

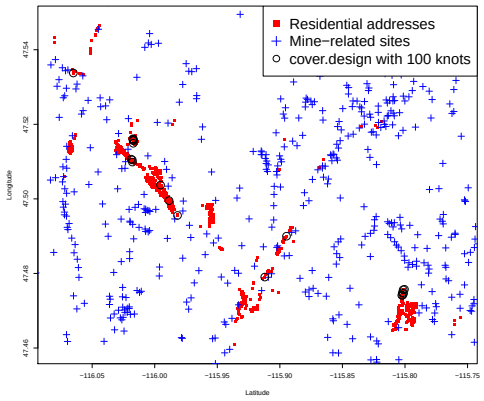
where $\boldsymbol{X}\boldsymbol{\beta}$ is a low-order polynomial (the *fixed effect*), and $\boldsymbol{Z}\boldsymbol{\alpha}$ is a non-linear function represented as *random effects* with covariance matrix $\boldsymbol{G}$ for the random effect $\boldsymbol{\alpha}$. The error term $\boldsymbol{\epsilon}$ is assumed to be independent as in model in Eq. (3.1).

There are number of alternatives for model defining $\boldsymbol{Z}$ in Eq. (3.2). Kammann and Wand (2003) proposed the use of radial basis functions with generalized covariance matrices, where they used the term low-rank kriging (for a more extensive presentation the reader should review Ruppert et al. (2003)). Low-rank kriging utilizes a reduced number of knots locations placed over the whole study area to define the spline functions $\phi_j(s)$. The idea is to assume that the spatial information available from the entire set of observed locations can be summarized in terms of a smaller but representative sets of locations, or knots.

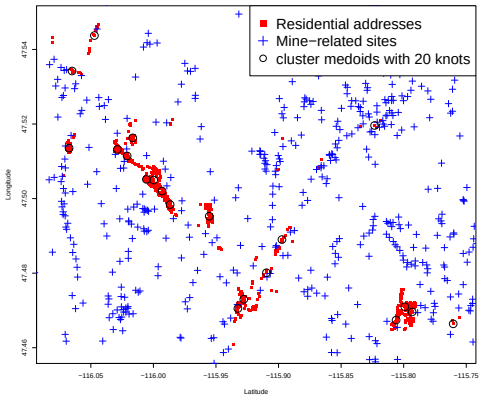
The spatial function is represented as a random effects term, $\boldsymbol{Z}\boldsymbol{\alpha}$, the variance of the random effects serves to penalize complex functions. Kammann and Wand (2003) suggest that $\text{Cov}(\boldsymbol{Z}\boldsymbol{\alpha}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$ is a reasonable approximation of the spatial covariance structure of the random effects. The classic geostatistical approach is based on a predefined chosen covariance function with corresponding parameters estimated *a priori* from a variogram analysis or likelihood methods (Diggle et al., 1998). The use of the variogram may be misleading in some situations (Diggle and Pinheiro, 2007) or when some of the implicit assumptions of kriging are violated or questionable. For the low-rank kriging approach, Wand (2003) proposes to construct $\boldsymbol{Z}$ based on the Matérn covariance. This method requires the selection of a smoothness parameter and a spatial range parameter that controls the smoothness of the fitted surface. The spatial range parameter is fixed to simplify the parameter estimation (French and Wand, 2004). In general the



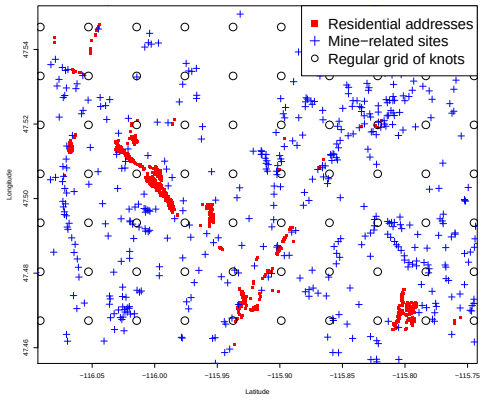
(a) Space-filling algorithm with 20 knots



(b) Space-filling algorithm with 100 knots



(c) Selection of 20 knots based on clustering algorithm



(d) Regular grid of 10×10 knots

Figure 3.2: Selection of

selection of the number and position of the knots is a complex optimization problem (Ruppert, 2000). The number of knots is determined by a simple rule

$$n.knots = \max\{20, \min(n/4), 150\}.$$

For the particular case of spatial smoothing, the selection of the locations of the knots is usually done by a geometric space-filling design based on a maximal separation principle (Johnson et al., 1990; Nychka and Saltzman, 1998) and implemented in the function `cover.design` available in the `R` package `fields`. Other options are to use a cluster technique and use the medoids locations as knots or use a regular grid. Hence the spatial structure is done through a dimension reduction based on the knots to define the spatial covariance function. Figure 3.2 illustrates the different alternatives for knots selection for the area of study. The locations of the residential addresses and mine-related sites are plotted and three different methods are shown: figures 3.2a and 3.2b show 20 and 100 knots chosen using the `cover.design` function in `fields` R package. Figure 3.2c shows 20 knots using a clustering algorithm related to the k-means algorithm (k-medoids algorithms) partitioning the locations into k clusters (Kaufman and Rousseeuw, 1987). In this case, each cluster corresponds to one knot location. The

effect of knots specificacion in two-dimensional data has not been investigated in depth. Kim et al. (2010) performed a sensitivity analysis for the selection of the number and location of the knots and compared the results with the full-rank kriging. They suggest that the results can be very sensitive to the choice of the spatial parameters (if it is choosen to be fixed as suggested in French and Wand (2004)). However, the use of low-rank kriging models are very sensitive to the selection of the number and position of the knots, with few knots the separation between them increases and the estimation of the spatial dependence and parameters become difficult (Ruppert et al., 2003; Kim et al., 2010). For the lead concentration levels, we found that the existence of high variability within a few kilometers or even within the same residential property caused difficulties for variogram analysis and the choice of an appropriate covariance structure for the selection of a spatial correlation.

We consider a more flexible approach a moderate large number of knots over a regular grid (as shown in figure 3.2d). The combination of tensor products of B-spline basis functions with penalties (commonly known as penalized splines or P-splines) are an attractive alternative for multidimensional smoothing (Eilers and Marx, 2003; Currie et al., 2006; Eilers et al., 2006; Lee and Durbán, 2010) commonly known as penalized splines or P-splines. B-spline basis functions (de Boor, 1978) and tensor products allow for good approximation of bivariate surfaces, although it can be extended to any number of covariates (see Wood, 2006a, Chapter 4). To illustrate the idea we consider two covariates x and z. Then for each covariate we represent a smooth function f_x and f_z that we write as:

$$f_x(x) = \sum_{k=1}^{K} \alpha_k \phi_k(x), \quad \text{and} \quad f_z(z) = \sum_{l=1}^{L} \beta_l \breve{\phi}_l(z),$$

where α_k and β_l are coefficients, and ϕ_k, and $\breve{\phi}_l$ are known basis functions. Let $A = [\alpha_{kl}]$ be a $K \times L$ matrix of coefficients, the bivariate surface is the represented as

$$f_{xz}(x, z) = \sum_{k=1}^{K} \sum_{l=1}^{L} \alpha_{kl} \phi_k(x) \breve{\phi}_l(z),$$

and so A may be chosen by least squares by minimizing

$$S = \sum_{i}^{n} \|y_i - f_{xz}(x, z)\|^2 = \sum_{i}^{n} \|y_i - \sum_{k=1}^{K} \sum_{l=1}^{L} \alpha_{kl} \phi_k(x) \breve{\phi}_l(z)\|^2, \tag{3.3}$$

where $\| \cdot \|^2$ denotes the $L2$-norm.

Penalized spline solution introduce a penalty function to the least squares problem in Eq. (3.3), defined as:

$$Pen(A) = \lambda_x \sum_{k} \|D_k \alpha_{k\bullet}\|^2 + \lambda_z \sum_{l} \|D_l \alpha_{\bullet l}\|^2, \tag{3.4}$$

where D_k and D_l are difference matrices of order q. Usually we choose $q = 2$, a quadratic or second order penalty, such that the difference matrix has the form:

$$D_k = \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & & & 0 \\ 0 & 1 & -2 & 1 & & & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & & & \\ & & & 1 & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ 0 & 0 & \cdots & & 0 & 1 & -2 & 1 \end{pmatrix}_{(k-q)\times k}, \tag{3.5}$$

and the same for D_l.

The first term of Eq. (3.4) puts a difference penalty on each column of A (i.e. $\alpha_{\bullet l}$) and the second term puts a difference penalties on each row of A (i.e. $\alpha_{k\bullet}$). Note that, λ_x and λ_z are smoothing parameters to control the amount of smoothing along the two dimensions, such that $0 < \lambda_x, \lambda_z < \infty$. An extreme example would be λ_x and $\lambda_z = \infty$ corresponding to polynomial regression (of order $q-1$) in the x-direction (where q is the penalty order), and a very light smoothing along the z-direction. We choose the $\phi(\cdot)$ as B-spline basis functions. B-spline basis functions are a very stable basis for large data (de Boor, 1978), and for spatial smoothing (Lee and Durbán, 2009). In compact form, the smooth function can be written as

$$f_{x,z}(x, z) = \boldsymbol{B}\boldsymbol{a},$$

where $\boldsymbol{a}$ is the vector of coefficients of length $KL \times 1$ and $\boldsymbol{B}$ is the tensor product of the two marginal B-spline bases $B_x = \phi_k(x)$ and $B_z = \phi_l(z)$, i.e.

$$\boldsymbol{B} = B_x \square B_z = (B_x \otimes 1_n') \odot (1_n' \otimes B_z), \quad \text{of dimension } n \times KL \tag{3.6}$$

where $\odot$ is the element-by-element or Hadamard product and $\otimes$ the Kronecker product. The combination of both matrix products with vectors of ones of length n as expressed in Eq. (3.6) is denoted by the row-tensor product by symbol $\square$ defined by Eilers et al. (2006). Figure 3.3 shows a sub-set of a tensor product of B-splines.

The solution for the basis coefficients is

$$\hat{\boldsymbol{a}} = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y}, \tag{3.7}$$

where $\boldsymbol{P}$ denotes the penalty on Eq. (3.4), in matrix form which is a kronecker sum:

$$\boldsymbol{P} = \lambda_x D_x' D_x \otimes I_K + \lambda_z I_L \otimes D_z' D_z, \tag{3.8}$$

where I_K and I_L are identity matrices of sizes K and L, respectively. The details of these methods are described by (Eilers et al., 2006), Wood (2006a) and many other authors. In particular, Lee and Durbán (2010) discuss P-splines in the spatial and spatio-temporal setting.

In practice, there are some parameters to be chosen: (i) the number of segments in which we divide the range of x and z (say $nseg_x$ and $nseg_z$ and where we define a set of equally spaced knots to make a regular grid), (ii) the order of the B-spline (usually cubic splines), (iii) and the order of the penalty in each dimension (usually second order). Then with cubic splines and second order penalties the size of each marginal B-spline basis is $n \times K$ and $n \times L$ respectively, where K are $nseg_x + 3$ and L is $nseg_z + 3$. Finally, the size of the regression matrix $\boldsymbol{B}$ is $n \times c$, where $c = KL$ is the length of the vector of coefficients $\boldsymbol{a}$ (see Eilers and Marx, 1996, for details).

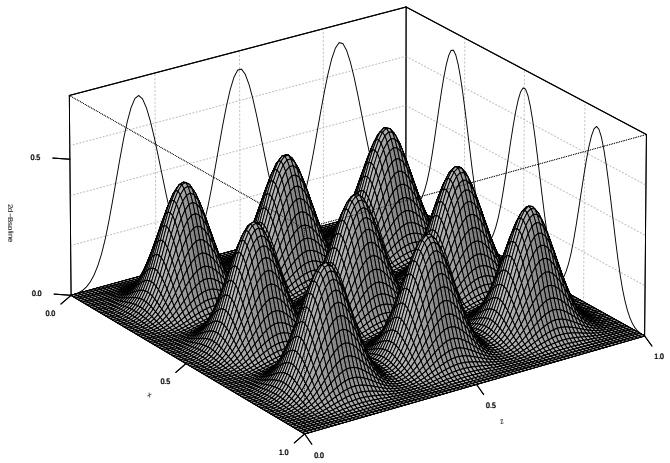


Figure 3.3: Portion of a 3×3 tensor product B-spline basis.

The computational advantage of using tensor products splines over kriging depends strongly on the number of basis functions. In almost all practical applications, a number of 25 basis functions for each dimension of the bivariate model over a regular grid of knots covering the region of study presents little computational challenge. The use of a second-order smoothness penalty encourages the appearance of linear sections if there is a gap in the data. In all forms of flexible regression or smoothing techniques, the choice of the degree of smoothness for the estimator is crucial. In the context of bivariate P-splines, we need to choose λ_x and λ_z. Most widely used approaches include cross-validation (CV), generalized cross-validation (GCV) or information criteria as a balance between the goodness-of-fit of the model against complexity, i.e. Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC). The details of selection criteria are discussed by many authors, with Wood (2006a) a good starting point.

The extension of the P-spline model as a mixed model approach as in Eq. (3.2) can be easily considered by the reparameterization of the model bases and coefficients. In general, this can be achieved in several ways as in (Eilers, 1999). Welham et al. (2007) give a comprehensive review of mixed model representations of spline models. In general, a computationally efficient method to reparameterize the model is the use of the singular value decomposition of the penalty matrix $D'D$ in one dimension, and similarly for the bivariate case to the simulatenous decomposition of the kronecker sum in Eq. (3.8) (see Lee and Durbán, 2010; Wood, 2006a, for details). The main advantage of the mixed model approach is the estimation of the amount of smoothing as a ratio of variances, and hence estimation and inference can be done using standard mixed model approaches as restricted maximum likelihood (Ruppert et al., 2003). These methods can be easily implemented in the statistical software R, with the function `gamm` in library `mgcv` (Wood, 2006b) and tensor product smooths with the function `te` (Wood, 2006b, 2011).

3.2 Bivariate Density estimation of mine-related sites

As showed in Figure 2.1 residential properties are surrounded by a variety of mine-related sites. In some cases, the residential properties exposed to heavy metal contamination might be due to proximity to a mine-related site. Therefore it is of interest to include this information in our analysis so as to account for it. Hence, we estimated the spatial density of mine-related sites in the area, and predicted the density for each of the residential properties.

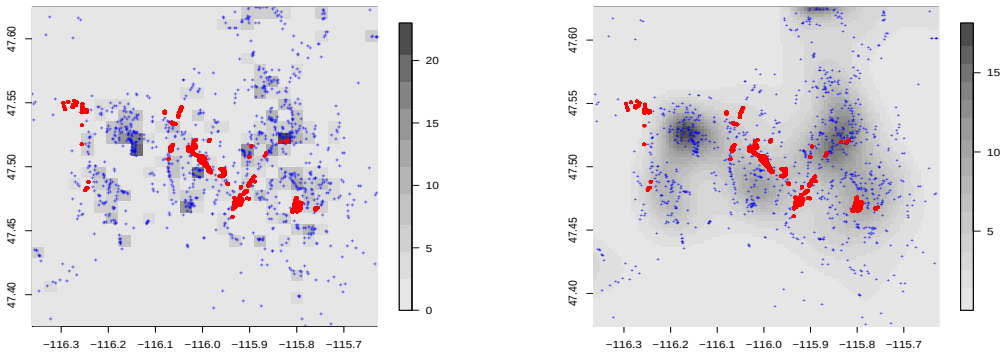
The density function can be estimated using different approaches, in fact it can be viewed as the estimation of the intensity function in spatial point patterns (Diggle, 1983). However, we do not assume any stochastic underlying point-process, as we only include the information of these sites as an additional covariate in the final model. In order to maintain an unified approach, we use tensor products splines instead of other techniques such as kernel density estimators. The bivariate tensor product splines provides a simple and effective density estimation approach (Eilers and Marx, 2006; Durban et al., 2006). The approach consists of pre-processing the data into a bivariate histogram and count the number of observation on each bin, then assume the data are Poisson counts and estimate the density as a penalized Poisson regression generalized linear model with a log link function (Nelder and Wedderburn, 1972). Figure 3.4a shows the bivariate histogram for the mine-related sites with 20 bins in each dimension, the residential properties in the study are also plotted. Figure 3.4b shows the smoothed density of mine-related sites, there is very little difference in the density fit if we use a different number of bins in the construction of the bivariate histogram as long as they are large enough. One of the advantages of this approach is the selection of the amount of smoothing, where we use an anisotropic density smooth with tensor products and B-spline bases implemented in the function `gamm` in the library `mgcv`. The estimation of the tensor product smooth models were implemented using `mgcv 1.7-24` in the software R release 3.0.1 (**?**). The tensor product smooths were constructed based on a 10×10 regular grid of knots over the region of study. The estimation of the Poisson regression model is performed using penalized quasi-likelihood (Breslow and Clayton, 1993). From Figure 3.4b we can see that some residential properties may be more exposed to heavy metals contamination due to proximity to an area with dense mine-related sites. The estimation of this density allows us to incorporate more spatial information to understand the spatial variation in the lead concentration levels in residential soil. In the next section, we incorporate these estimates as a covariate in the spatial model. Hence, we are implicitly assuming a relationship between the density of the mine-related sites surrounding the property and the concentration levels of lead in residential soil.

3.3 Geoadditive modelling of lead concentration levels

We use a smooth model to describe the spatial variability of lead in residential soil. In order to reduce the data skewness we consider the logarithm of Pb. The model is defined as:

$$\log(\widehat{\mathrm{Pb}})_{ijk} = \gamma_0 + \gamma_{1j} + \gamma_{2k} + f(\texttt{Location}_i) + s(\texttt{Density}_i), \tag{3.9}$$

where $\log(\widehat{\mathrm{Pb}})_{\mathrm{ijk}}$ is the $\log$ of concentration level at residential property i, sampled at j^{th} location and k^{th} depth, γ_0 is an intercept term, and γ_{1j}, and γ_{2k} are the coefficients for the factor variables `SAM_LOC` ($j = 1, ..., 8$) for sample location and `LAYER` ($k = 1, 2, 3, 4$)



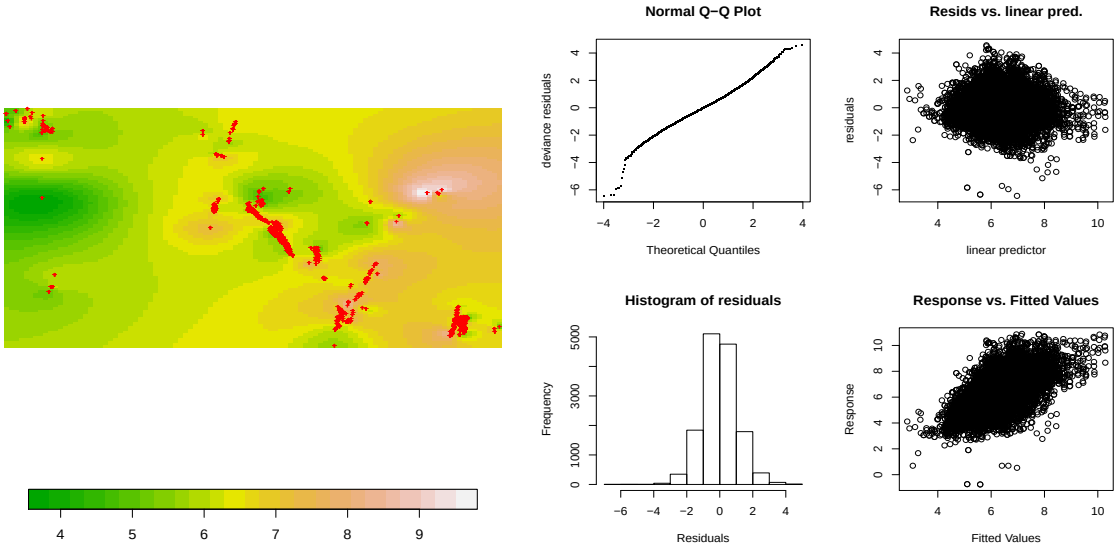
(a) Bivariate histogram

(b) Bivariate density estimation

Figure 3.4: Residential properties (red squares) and mine-related sites (blue crosses). Left: bivariate histogram. Right: Smoothed density of mine-related sites

for sample depth respectively. The function $f(\cdot)$ is a bivariate P-spline tensor product smooth of the `Location` for each residential property in terms of the geographical coordinates as shown in section 3.1. For each dimension we considered the same set of 10×10 regular grid knots defined in Section 3.2 to estimate the bivariate density. The function $s(\cdot)$ is a univariate P-spline smoother of the predicted density, `Density`, at each i^{th} location estimated in Section 3.2. The advantage of estimating the density as a poisson regression model is that we can estimate the density at new locations using the regression function. Hence, we are including the predicted density as an additional covariate in model in Eq. (3.9), and therefore we are intrinsically assuming that there is possibly a non-linear relationship between log Pb mean concentration levels in residential soil and the density of mine tailings surrounding the property. We also assume that the residuals are i.i.d. Gaussian.

The fitted spatial surface, $f(\texttt{Location}_i)$, of model in Eq. (3.9) for Pb is shown in Figure 3.5a. Note that we interpolate the estimated surface over a rectangular region in order to allow us to visualize the spatial distribution of the log concentrations of Pb in the whole area. Some residuals checking plots are shown in Figure 3.5b. These plots show that the Gaussian assumption should be carefully considered due to the existence of extreme outliers. The effect of the sample location and depths are shown in Table 3.1 and Figure 3.6 shows the partial effects for comparison. From Table 3.1, we find that there are significant differences between all the sample locations and the Driveway sample (except for the Right-of-Way location). Standard errors are large for some levels due to the high variability and the small number of samples for those sample locations (Garage, and Play area samples) as shown in Table 3.1. For sample depths, it can be noticed that for log Pb concentrations at A(0-1 in) and B(1-6 in) depths are not significantly different, and also that the deepest sample intervals (i.e. C(6-12 in) and D(12-18 in)) have lower Pb concentrations. The results shows that soil samples located in driveways, parking and Right-of-Way locations have higher levels than those samples located in the garden, play area or yard. This result suggest that lead and heavy metals in general may be transported through roads as dust. However, these results must be considered with some caution. Remedial actions were taken in past years through clean-up activities in



(a) Estimated spatial surfaces for $\log(\mathrm{Pb})$ (b) Residual plots of estimated models for $\log(\mathrm{Pb})$.

Figure 3.5: Estimated surface and residuals plots for $\log(\mathrm{Pb})$

some residential properties. The residential remedial program effectively replaced contaminated surface soils in specific areas such as yards and play areas where children are more exposed to heavy metals contamination (**?**). The information regarding which residential properties were cleaned and remediated in the previous years were not available for this study.

There are different alternatives to tackle the possible violation of assumptions evidenced in the residual plots: (i) consider more flexible distributions (e.g. Gamma with log-link), (ii) consider generalized additive models for location, shape and scale with distributions for skewed data (GAMLSS, see Rigby and Stasinopoulos (2005)), or (iii) other transformations on the data to achieve more symmetry and maintain the Gaussian assumption. Alternatively, given that our aim is to analyze the spatial distribution of the data, we consider a simpler approach commonly used in the analysis of geochemical samples. We grouped the $\log(\mathrm{Pb})$ values of those observations with the same sample location and depth levels and computed the geometric mean, $\mathrm{Pb}_{\mathrm{gm}}$ (i.e. samples in the same location and measured at the same depth in a residential property are averaged using the geometric mean, then sample location and depth levels are averaged for each residential property). With the geometric mean the effect of the outliers is dampened, and gives a unique representative measure of the Pb concentration levels for each residential property. The model for the $\log$ mean concentration levels of Pb is:

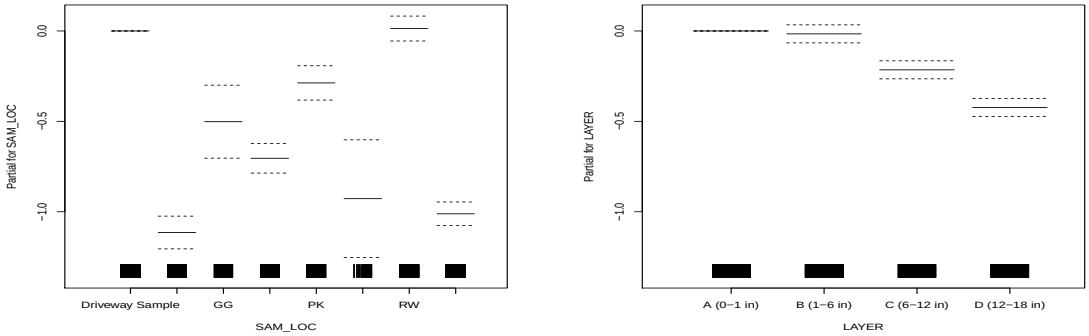
$$\log(\widehat{\mathrm{Pb}}_{\mathrm{gm}})_i = \gamma_0 + f(\texttt{Location}_i) + s(\texttt{Density}_i). \tag{3.10}$$

The fitted surface and residuals plots for model in Equation (3.10) are shown in Figure 3.7. The fitted spatial surface does not differ much compared to the estimated surface for model in Equation (3.9), but residuals seem to be more adequate based on Gaussian error assumptions.

Figure 3.8 shows the estimated smooth effects for the density of mine-related sites for $\log(\mathrm{Pb})$ and $\log(\mathrm{Pb}_{\mathrm{gm}})$. In both cases, the effect of the smoothed density of mine-related

Table 3.1: Estimate intercept, sample location and depth coefficients

$\log(\text{Pb})$	Coefficient	Std.Error	p-value
(Intercept)	7.15	0.03	0.00
Garden Sample	-1.11	0.05	0.00
Garage	-0.50	0.10	0.00
Other Sample	-0.74	0.04	0.00
Parking	-0.29	0.05	0.00
Play Area Sample	-0.93	0.16	0.00
Right-of-Way	0.01	0.03	0.70
Yard Sample	-1.01	0.03	0.00
B (1-6 in)	-0.02	0.03	0.52
C (6-12 in)	-0.21	0.03	0.00
D (12-18 in)	-0.42	0.03	0.00



(a) sample location effect for $\log(\text{Pb})$ (b) sample depth effect for $\log(\text{Pb})$

Figure 3.6: Partial effects for sample location and sample depth.

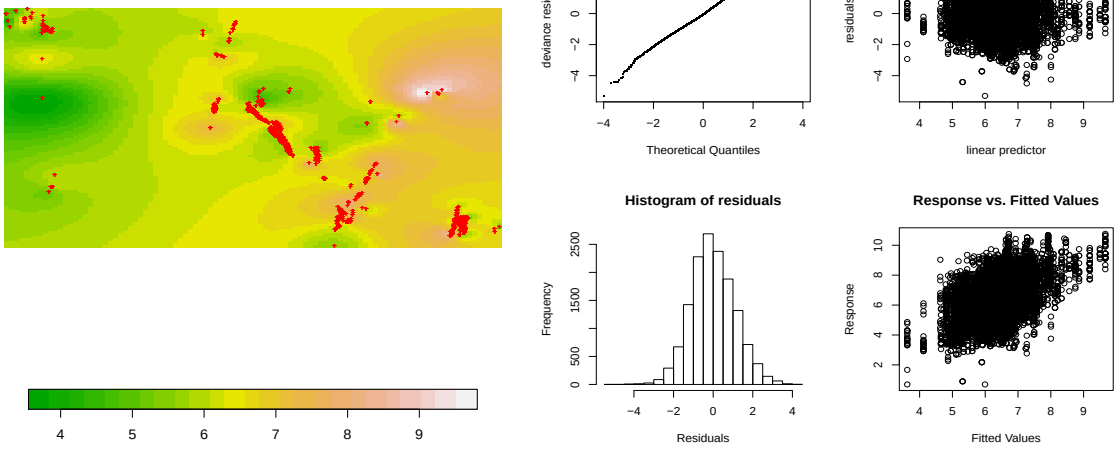
sites is very similar, and the interpretation of this effect is straightforward: high density mine-related sites contribute to increased the Pb concentration levels in residential soil.

4 Geostatistical risk assessment of lead concentration in the Coeur D'Alene River Basin

In this section we estimate possible risks of adverse health outcomes, providing a geostatistical analysis of high-risk residential properties. We fitted a spatial logistic model where the outcome is a Bernoulli response indicating if the Pb concentration level is greater that the established action level of 1000 mg/kg for Pb. If lead concentration exceeds 1000 mg/kg, contaminated soil is partially removed (to the appropriate depth) and replaced with clean soil, defined as containing less than 100 mg/kg of Pb.

The general formulation for a spatial logistic regression is:

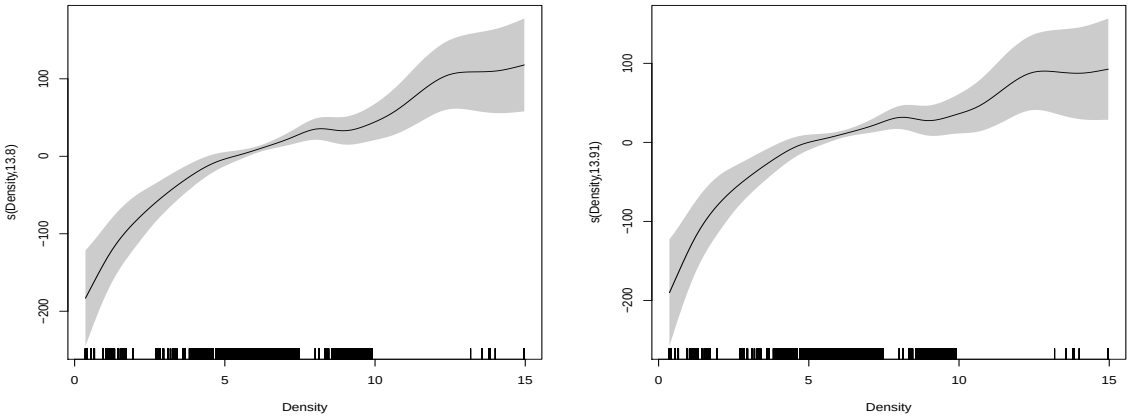
$$\begin{aligned} z_i &\sim \text{Bern}(p(\boldsymbol{x}_i, \boldsymbol{s}_i)) \\ \text{logit}(p(\boldsymbol{x}_i, \boldsymbol{s}_i)) &= g(\boldsymbol{x}_i, \boldsymbol{s}_i), \end{aligned} \tag{4.1}$$



(a) Estimated spatial surfaces for $\log(\mathrm{Pb}_{\mathrm{gm}})$

(b) Residual plots of estimated models for $\log(\mathrm{Pb}_{\mathrm{gm}})$.

Figure 3.7: Estimated surface and residuals plots for $\log(\mathrm{Pb}_{\mathrm{gm}})$



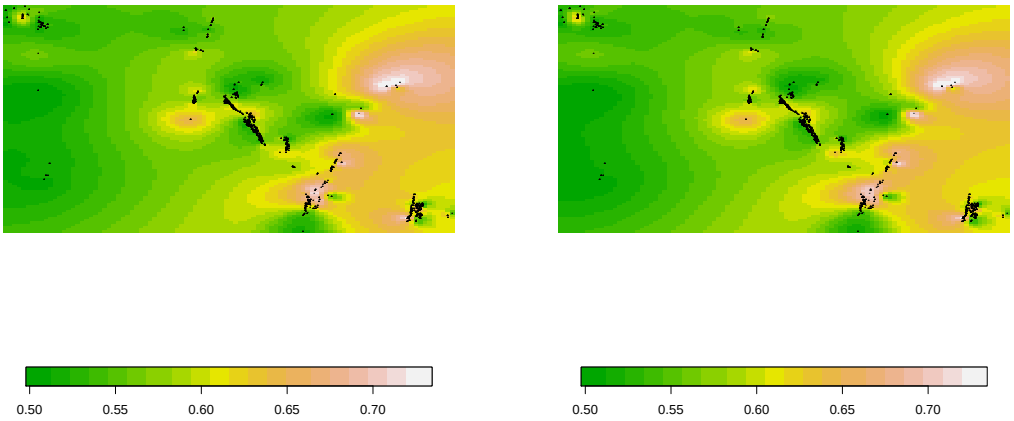
(a) Smoothed density effect for $\log(\mathrm{Pb})$

(b) Smoothed density effect for $\log(\mathrm{Pb}_{\mathrm{gm}})$

Figure 3.8: Mine-related density effects for Pb and $\mathrm{Pb}_{\mathrm{gm}}$ concentration levels in residential soil.

where z_i is the binary data indicating if the sampled value exceeds the threshold action level (1000 mg/kg), $\boldsymbol{x}_i$ is a vector of covariates, $\boldsymbol{s}_i$ denotes the spatial locations, and $g(\cdot)$ is a function of the $\boldsymbol{x}_i$ covariates and the spatial locations $\boldsymbol{s}_i$. Model in Equation (4.1) is a common approach in spatial epidemiology for the estimation of disease risk factors (Prentice and Pyke, 1979; Elliot et al., 2000). We use a logit link for assessing the relative risk based on the covariates, and penalized quasi-likelihood for estimation.

Now, we estimate the spatial logistic regression models for the binary response:



(a) Smoothed spatial risk surfaces for z_i (b) Smoothed spatial risk surfaces for z_i^{gm}

Figure 4.1: Spatial risk for z_i and z_i^{gm}

$$z_i = \begin{cases} 1 & \text{if Pb}_i > 1000\ mg/kg \\ 0 & \text{if Pb}_i < 1000\ mg/kg \end{cases} \quad \text{and } z_i^{\text{gm}} = \begin{cases} 1 & \text{if Pb}_i^{\text{gm}} > 1000\ mg/kg \\ 0 & \text{if Pb}_i^{\text{gm}} < 1000\ mg/kg \end{cases}$$

where z_i is calculated from the values sampled by location and depth, and z_i^{gm} from the geometric mean computed by sample location and depth values at each residential address. Then, for the $z_i \sim \text{Bern}(p(\boldsymbol{x}_i, \boldsymbol{s}_i))$, we have that $\text{logit}(p(\boldsymbol{x}_i, \boldsymbol{s}_i))$ in model (4.1) becomes:

$$\text{logit}(p(\boldsymbol{x}_i, \boldsymbol{s}_i)) = \gamma_0 + \gamma_{1j} + \gamma_{2k} + f(\texttt{Location}_i) + s(\texttt{Density}_i), \tag{4.2}$$

and for $z_i^{\text{gm}} \sim \text{Bern}(p(\boldsymbol{x}_i, \boldsymbol{s}_i))$:

$$\text{logit}(p(\boldsymbol{x}_i, \boldsymbol{s}_i)) = \gamma_0 + f(\texttt{Location}_i) + s(\texttt{Density}_i). \tag{4.3}$$

For both models $f(\texttt{Location}_i)$ and $s(\texttt{Density}_i)$ are the smooth functions for the spatial surface and for the density of mine-related sites, respectively, as discussed in Section 3.3. Note that in the previous section we aimed to model the lead concentration levels, now we are interested in the estimating a risk measure (the probability of an individual sample exceeding the action level) such that remediation would be required. Using the unified approach for modelling Pb levels and Pb geographical risk, now we reformulate the problem into a generalized linear model for binary data. The essence of the spatial surface estimation by tensor products remains the same. Comparisons of different alternative approaches for spatial logistic regression models as in Eq. (4.1) are investigated in Paciorek (2007).

Figure 4.1 show the predicted risk (probability of exceeding 1000 mg/kg Pb levels) surfaces based on models in Eq. (4.2) and Eq. (4.3). Both surfaces are very similar, highlighting those areas with higher risk of exceedance. However, model (4.2) allows us to predict the probability of exceeding the action level for each sample location and depth

levels, whereas model in Eq. (4.3) gives us the probability that the geometric mean for each residential property exceeds 1000 mg/kg of Pb. Indeed, the use of the geometric mean for the lead concentration levels gives a reasonable measure of the risk associated for an individual residence, thus helping to identify possible residential addresses for remedial action.

The estimation of the density of mine-sites is also similar (not shown) for both models, thus having the same effects as shown in Figure 3.8, although the confidence bands are wider, but this is a known problem in spatial models for binary outcomes given that the data contains much less information than continuous observations. For model in Equation (4.2), the sample location and depth parameters coefficients follow similar patterns as in Figure 3.6, i.e. higher risk levels area associated with driveways and Right-of-Way locations, and lower levels for garden, play area and yard samples. For sample depths, A(0-1 in) and B(1-6 in) samples have higher probabilities of exceeding the action level.

5 Conclusion

We have performed an analysis of the spatial distribution of lead concentration from a sample of residential properties in the Coeur D'Alene river basin area. We adopted penalized regression splines with tensor product smooths to undertake the analysis. This approach gives us a surface that characterizes the spatial distribution over the study region. The aim of the paper was not to compare alternative spatial methods, but to provide a flexible methodology, that is a good compromise between quality of fit, and interpretability of the spatial process. None of the previous analysis of heavy metal concentration levels in residential addresses in the CDRB have performed a geostatistical analysis of the data. In fact, the survey sampling strategy was performed with no statistical or spatial design. This paper presents a retrospective analysis of the collected data.

There are a number of possibilities for analysis of this type of data, such as Gaussian Markov Random Fields (Cressie, 1993; Stein, 1999; Banerjee et al., 2004; Rue and Held, 2005), and Bayesian techniques (Rue et al., 2009). In this paper, we consider tensor products of B-spline basis as a building block and for simplicity, and no model comparisons were performed. We consider that for more complex models, hierarchical Bayesian approaches are a very powerful tool for spatial data smoothing and in particular for geographical risk assessment. In fact, mixed models are connected to hierarchical Bayesian models, and hence, the implementation of the methodology presented in this paper with tensor product smooths in a Bayesian context can be easily implemented using Win/OpenBUGS (Crainiceanu et al., 2005; Lunn et al., 2009).

The survey samples considered in this paper were not collected for spatial data analysis, but instead residential properties were targeted to those with children and pregnant women. Due to the high variability in the soil samples within the same residential property, we averaged the values using the geometric mean to group the Pb concentration levels and give a less variable measure of Pb concentration levels for each residential property. Additionally, incorporating the density of mine-related sites in the study region, helps to relate the level of Pb in residential properties with a measure of the proximity to a mine-related site. It should be notice that the geographical characteristics of the area, the presence of roads, streams, past flood events, may be unmeasured

covariates that may vary spatially and contribute to the spatial distribution of Pb concentration levels in the Coeur D'Alene river basin. Furthermore, the estimation of the risk of exceedance gives an initial model to highlight hot spots for geographically targeted intervention.

Recent advances in spatial survey sampling can benefit from the type of models proposed in this paper. Environmental agencies can use the spatial models in order to design the survey. In this paper, we showed how Pb concentration levels of residential property soil levels of Pb are related to the density of mine-related sites surrounding the area. Geostatistical risk models proposed in section 4 may be useful for spatial targetting survey designs, given the costs of environmental sampling of soil lead concentration level (sampling effort and time). Future work will aim to design optimal spatial sampling strategies for field work.

Inference and prediction for spatial data are affected substantially by the spatial configuration of the sampling locations where measurements are taken. Most of the geostatistical models implicitly assume that sampling locations and measurements values are independent. However, in practice it is usual to collect data points at locations where higher (or smaller) values than the average of the outcome are expected. Diggle et al. (2010) use the term *preferential* sampling when the spatial locations depend on the expected value of the measurement at that location, meaning that there is a stochastic dependence between the sampling locations and the outcome. For instance, given the effect of the density of mine-related sites, one may expect to sample in those residential properties with high density of mine-related sites or that may have potential risk given some prior knowledge. However, a sampling scheme with heavier monitoring around potentially high outcome values will have the effect of over-estimating the response variable levels over the entire area, while heavier monitoring around low value areas would produce under-estimates. Another approachs to explore for environmental survey sampling is to consider surveys designs based on model (4.2) and (4.3), where sample probabilities will be based on the predicted risk.

Acknowledgements

This work was funded by an NIH grant for the Superfund Metal Mixtures, Biomarkers and Neurodevelopment project 1PA2ES016454-01A2.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Monographs on Statistics and Applied Probability 101. Chapman & Hall/CRC.

Breslow, N. E. and Clayton, D. G. (1993). Aproximated inference in generalised linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Assoc.*, 93(443):961–994.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, 92(1):91–103.

Cressie, N. (1993). *Statistics for Spatial Data (Revised Edition)*. John Wiley and Sons, Inc., New York.

Currie, I. D., Durbán, M., and Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B*, 68:1–22.

de Boor, C. (1978). *A practical Guide to Splines*. Springer, Berlin.

Diggle, P. J., Menezes, R., and Su, T. (2010). Geostatistical inference under preferential sampling. *J. R. Statist. Soc. C (Applied Statitstics)*, 59:191–232.

Diggle, P. J. and Pinheiro, P. J. (2007). *Model-based geostatistics*. Springer.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, 47:299–350.

Durban, M., Currie, I. D., and Eilers, P. H. C. (2006). Mixed models, array methods and multidimensional density estimation. In *In Proceedings of the 21st International Workshop on Statistical Modelling*.

Eilers, P. H. C. (1999). Discussion of 'The analysis of designed experiments and longitudinal data by using smoothing splines' (by a. p. Verbyla, b. r. cullis, m. g. kenward, and s. j. welham). *Appl. Statist.*, 48:307–308.

Eilers, P. H. C., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, 50(1):61–76.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Sci.*, 11:89–121.

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66:159–174.

Eilers, P. H. C. and Marx, B. D. (2006). Multidimensional density smoothing with P-splines. In *In Proceedings of the 21st International Workshop on Statistical Modelling*.

Elias, R. W. and Gulson, B. (2003). Overview of lead remediation effectiveness. *The Science of the Total Environment*, 303:1–13.

Elliot, P., Wakefield, J., Best, N., and Briggs, D. (2000). *Spatial epidemiology: methods and applications.* Oxford University Press.

French, J. L. and Wand, M. P. (2004). Generalized additive models for cancer mapping with incomplete covariates. *Biostatistics*, 5(2):177–191.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Characterization*. Springer.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society, C - Applied Statistics*, 52:1–18.

Kaufman, L. and Rousseeuw, P. J. (1987). *Clustering by means of medoids*. Statistical Data Analysis based on the L-1 norm and related methods. Y. dodge, north-holland edition.

Kim, J., Lawson, A. B., McDermott, S., and Aelion, C. M. (2010). Bayesian spatial modeling of disease risk in relation to multivariate environmental risk fields. *Stati*, 29:142–157.

Laslett, G. M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications. *J. Am. Stat. Assoc.*, 89:391–409.

Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics and Data Analysis*, 53(8):2958–2979.

Lee, D.-J. and Durbán, M. (2010). *P*-spline ANOVA-type interaction models for spatio-temporal smoothing. *to appear in Statistical Modelling*.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Stat. Soc., B*, 61:381–400.

Lindern, I., S., S., Petroysan, V., and Von Braun, M. (2003). Assessing remedial effectiveness through the blood lead: soil/dust lead relationship at the Bunker Hill Superfund Site in the Silver Valley of Idaho. *The Science of the Total Environment*, 303:139–170.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The bugs project: Evolution, critique, and future direction. *Statistics in Medicine*, 28:3049–3067.

Nelder, J. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Stat. Soc. A*, 135:370–384.

Nychka, D. and Saltzman, N. (1998). *Design of air-quality designs.*, chapter In Case studies in Environmental Statistics. Lecture notes in Statistics. New York: Springer.

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large data sets. *Comput. Stat. Data An.*, 51:3631–3653.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–412.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, 54(3):507–554.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B*, 71:319–392.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. Technical report.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, UK. ISBN: 0521785162.

Spalinger, S. M., Von Braun, M. C., Petrosyan, V., and Von Lindern, I. H. (2007). Northern idaho house dust and soil lead levels comparted to the Bunker Hill Superfund Site. *Environ. Monit. Assess*, 130:57–72.

Stein, M. L. (1999). *Interpolating Spatial Data: Some Theory of Kriging*. Springer-Verlag, New York.

TerraGraphics (2003). Final quality assurance project plan (QAPP) for residential property sampling in the Coeur D'Alene River Basin of Idaho.

U.S. Environmental Protection Agency (2002). Record of Decision (ROD) - Bunker Hill Mining and Metallurgical Complex Operable Unit 3 (Coeur D'Alene Basin).

Verbyla, A., Cullis, B., Kenward, M., and Welham, S. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *J. Roy. Stat. Soc. C*, 48:269–312.

Wahba, G. (1990). Letter to the editor: Comment on cressie. *The American Statistician*, 44:255–256.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, 18:223–249.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *J. Am. Stat. Assoc.*, 93(441):341–348.

Welham, S. J., Cullis, B. R., Kenward, M. G., and Thompson, R. (2007). A comparison of mixed model splines for curve fitting. *Aust. N. Z. J. Stat.*, 49(1):1–23.

Wood, S. N. (2006a). *Generalized Additive Models - An introduction with R*. Texts in Statistical Science. Chapman & Hall.

Wood, S. N. (2006b). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Statist. Soc. B*, 73:3–36.